# Zero-Maintenance Disk Arrays
(Fast Abstract)

Jehan-François Pâris
Department of Computer Science
University of Houston
Houston, TX, USA
jfparis@uh.edu

Darrell D. E. Long[†]
Department of Computer Science
University of California
Santa Cruz, CA, USA
darrell@cs.ucsc.edu

Thomas Schwarz, S. J.
Departamento de ICC
Universidad Católica del Uruguay
Montevideo, Uruguay
tschwarz@ucu.edu.uy

## I. Introduction

Magnetic disks are the least reliable component of most computer systems. In addition, their failures result much more often in irrecoverable data losses than failures of all other components. As a result, nearly all medium to large-size disk arrays include some redundancy and provisions for the quick replacement of failed units.

These provisions are not difficult to implement in installations that have trained personnel on site. When this is not the case, each disk replacement will require a service call whose cost is likely to exceed that of the equipment being replaced. This is even truer for installations that are far away of metropolitan areas.

We are currently investigating the feasibility of building zero-maintenance disk arrays that free users from all maintenance tasks over the expected lifetime of the array. Three challenges are to be met to make these organizations cost-effective. These are the initial cost of the array, its performance and its long-term reliability.

Our preliminary results indicate that the key factor in the feasibility of zero-maintenance arrays is the failure rate of idle spares. In particular, cost effective solutions for archival data storage are possible as long as these failure rates remain negligible. For instance, a zero-maintenance disk array with 56 data disks, 8 parity disks and 50 spare disks would have a better than 99.999 percent probability of not experiencing a data loss over five years while having a space overhead comparable to that of a mirrored organization consisting of 56 pairs of disks.

This is less true if we assume that unused spare disks fail at the same rate as active disks. As the disk array ages, its supply of spare disks gets progressively depleted. As a result, the array will require 62 to 70 percent extra spares to achieve the same reliability level. For instance, a zero-maintenance disk array with 56 data disks and 8 parity disks would require 82 spare disks instead of 50 to achieve the same 99.999 percent probability of no data loss over five years. These extra spare disks result in a much higher space overhead, which then approaches that of an organization that would keep three copies of all its data.

In reality, we may expect unused spare disks to fail at a significantly lower rate than active disks. This will reduce the need for additional spare disks and make zero-maintenance disk arrays more attractive.

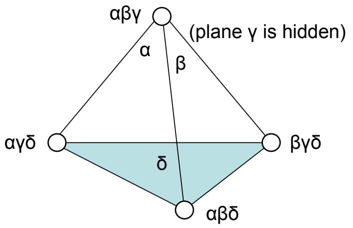

Fig. 1. A three-dimensional RAID organization using four parity disks to protect the contents of four data disks against all triple disk failures. Each data disk is identified by the three parity planes to which it belongs.

## II. Background

Sparing is a well-known technique for increasing the reliability of disk arrays. Adding a spare disk to an array provides the replacement disk for the first failure. Distributed sparing [1] gains performance benefits in the initial state and degrades to normal performance after the first disk failure.

Many of the data we store online are archival in nature and are not supposed to be modified once they are stored. As a result, write throughputs are much less important in archival storage systems than in conventional storage systems. On the other hand, archival file systems have fairly stringent reliability requirements as they have to guarantee the integrity of their data over periods of time that often measure in decades.

Three-dimensional RAID arrays [2] extend to three dimensions most of the properties of two-dimensional RAID arrays [3, 4, 5]. A three-dimensional RAID array consists of $n_p$ parity stripes, each containing a single parity disk. These parity stripes will form $n_p$ parity planes in a three-dimensional space. We assume that all these planes intersect with each other and that three arbitrary planes intersect at exactly one single point. We then place one data disk at each of these triple intersections for a total of $n_d = \binom{n_p}{3}$ data disks. Since each of these data disks belongs to three parity planes, it will be able to recover from all triple failures.

Fig. 1 represents a three-dimensional RAID array with four data disks and four parity planes As we can see, the four intersecting planes α, β, γ, and δ define a tetrahedron with six edges and four summits, and each of these summits defines a data disk, to which we attach the labels of the three intersecting parity planes. Observe that the array will be able to recover from all triple disk failures and most quadruple failures without any data loss, the sole exception being a failure of all four data disks αβγ, αβδ, αγδ and βγδ.

[†] Supported in part by Grant CCF-1219163, by the Department of Energy under Award Number DE-FC02-10ER26017/DE-SC0005417 and by the industrial members of the Storage Systems Research Center.

TABLE I. FIVE-YEAR SURVIVAL RATES AND SPACE OVERHEADS OF SELECTED ZERO-MAINTENANCE DISK ARRAYS ASSUMING THAT SPARE DISKS DO NOT FAIL UNTIL THEY ARE PUT TO USE.

| Number of | | | Space Overhead | Ninety-five percent C I for five-year reliability |
|---|---|---|---|---|
| Parity disks | Data disks | Spare disks | | |
| 5 | 10 | 16 | 0.678 | (5.3, 5.5) |
| 6 | 20 | 24 | 0.600 | (5.0, 5.2) |
| 7 | 35 | 35 | 0.545 | (5.0, 5.2) |
| 8 | 56 | 50 | 0.509 | (5.2, 5.4) |
| 9 | 84 | 66 | 0.472 | (5.0, 5.1) |

TABLE II. FIVE-YEAR SURVIVAL RATES AND SPACE OVERHEADS OF SELECTED ZERO-MAINTENANCE DISK ARRAYS ASSUMING THAT SPARE DISKS FAIL AT THE SAME RATE AS OTHER DISKS.

| Number of | | | Space Overhead | Ninety-five percent C I for five-year reliability |
|---|---|---|---|---|
| Parity disks | Data disks | Spare disks | | |
| 5 | 10 | 26 | 0.756 | (5.0, 5.2) |
| 6 | 20 | 41 | 0.701 | (5.2, 5.5) |
| 7 | 35 | 58 | 0.650 | (5.0, 5.2) |
| 8 | 56 | 82 | 0.616 | (5.1, 5.3) |
| 9 | 84 | 110 | 0.586 | (5.1, 5.3) |

## III. ZERO-MAINTENANCE ARRAYS

Our main objective was to achieve 99.999 percent reliability over a five-year interval while keeping the space overhead as low as possible. We selected three-dimensional RAID arrays as our base organization because of their high reliability and their low space overhead and added enough spare disks to ensure that the array had a 99.999 percent chance of not losing any data over five years. This combination proved itself much more successful than a previous attempt [6].

## IV. RELIABILITY ANALYSIS

We evaluated the reliability of zero-maintenance disk arrays under two distinct hypotheses, namely, that spare disks would not fail until they are put to use and that spare disks would always fail at the same rate as the other disks.

Tables I and II summarize the results of our analysis. These results were obtained using two modified versions of the Proteus disk array simulator [7] that corresponded to our two hypotheses. The disks arrays were assumed to tolerate all triple failures of their data disks and their parity disks and most quadruple failures [2]. All quintuple failures were assumed to be fatal. Reliability values are expressed in "nines." Four nines correspond to 99.99 percent reliability and five nines to a 99.999 percent value.

We assumed that disk failures were independent events distributed according to a Poisson law with a mean time to failure (MTTF) of 100,000 hours. This value is at low end of the values observed by both Schroeder and Gibson [8] and Pinheiro, Weber and Barroso [9]. Repair times were assumed to be deterministic and equal to 12 hours. Space overhead ratios were computed by dividing the number of disks that did not hold data (parity disks and spare disks) by the total number of disks in the array.

Both tables show that the space overhead that is needed to obtain 99.999 percent reliability decreases as the size of the array increases. Two factors explain that. First, the parity disk to data disk ratio decreases rapidly as the size of the array increases. Second, the central limit theorem predicts that the coefficient of variation of the number of disk failures that will occur over a five year interval will decrease proportionally to the square root of the size of the array.

Looking at the results in Table I, which assume that unused disks will not fail until they are put to use, we can see that zero-maintenance disk arrays with at least 20 data disks require a space overhead of no more than 60 percent to achieve a five-year reliability of five nines (99.999 percent). As the sizes of the arrays increase, their space overheads become comparable to those of arrays of mirrored disks.

As Table II shows, the same is not true when we consider that unused spare disks will fail at the same rate as the other disks. As disk arrays age, their supply of spare disks will now get progressively depleted. As a result, the arrays require 62 to 70 percent extra spares to achieve the same five-year reliability level.

One of the most promising approaches for reducing this overhead would be to monitor in real time the current number of available spare disks and notify the manufacturer whenever this number dips below a critical threshold. The occurrence would then be handled by all parties as a regular warranty repair.

## V. CONCLUSION

We have presented a novel architecture for disk arrays that does not require users to perform any maintenance tasks over the expected lifetime of the array. Preliminary results indicate that the key factor in the feasibility of our design is the failure rate of unused spare disks. As long as these rates remain negligible, zero maintenance disk arrays with at least 77 disks can provide a five-year reliability of five nines (99.999 percent) with a space overhead comparable to that of mirroring. If this is not the case, we would need between 64 and 70 percent extra spare disks to achieve the same five-year reliability, which would result in a higher space overhead.

## REFERENCES

[1] A. Thomasian and J. Menon. "RAID 5 performance with distributed sparing." IEEE Trans. on Parallel and Distributed Systems, 8(6):640–657, June 1997.

[2] J.-F. Pâris, D. D. E. Long and W. Litwin, "Three-dimensional redundancy codes for archival storage," Proc. 21st Int. MASCOTS Symp., Aug. 2013.

[3] L. Hellerstein, G. Gibson, R. M. Karp, R. H. Katz, and D.A. Patterson, "Coding techniques for handling failures in large disk arrays," Algorithmica, 12(3-4):182-208, June 1994.

[4] T. J. E. Schwarz, "Reliability and Performance of Disk Arrays," Ph.D. Thesis, Dept. CSE, U. C. San Diego, 1994.

[5] J.-F. Pâris, A. Amer, and T. J. E. Schwarz, "Low-redundancy two-dimensional RAID arrays," Proc. 2012 Int. ICNC Conf. pp. 507–511, Jan.-Feb. 2012.

[6] J.-F. Pâris and T. J. E. Schwarz. "On the possibility of small, service-free disk based storage systems," Proc. 3rd Int. ARES Conf., pp. 56–63, Mar. 2008.

[7] H.-W. Kao, J.-F. Pâris, T. Schwarz, SJ and Darrell D. E. Long, "A flexible simulation tool for estimating data loss risks in storage arrays," *Proc 29th IEEE MSST Conf.*, May 2013.

[8] B. Schroeder and G. A. Gibson, "Disk failures in the real world: what does an MTTF of 1,000,000 hours mean to you?" Proc. 5th USENIX FAST Conf., pp. 1–16, Feb. 2007.

[9] E. Pinheiro, W.-D. Weber and L. A. Barroso, "Failure trends in a large disk drive population," Proc. 5th USENIX FAST Conf., pp. 17–28, Feb. 2007.